利用 URL-Key 领域术语识别方法

吕书宁[1] 董志安[2†]

1. 北京工业大学软件学院，北京 100124 ；2.北京信息科技大学网络文化与数字传播北京市重点实验室，北京 100101：†通讯作者，dong.zhian@163.com

摘要 本文首次提出了利用 URL-Key 进行领域术语识别方法。以 URL 作为媒介，借助已知 URL-Key 的领域性来判断未知领域候选数据的领域性。首先借助互联网中已有的人工分类的领域 URL,根据 URL-Key 在各领域汇总使用的频度，提出基于方差的领域 URL-Key 的识别方法，构建领域 URL-Key 词表。然后利用伪反馈技术，收集候选领域词检索得到的 URL 结果集，根据 URL 结果集构建候选领域术语的 URL-key 特征向量，最后利用 SVM 对候选领域术语进行提取。本文在 4 个领域上对实验加以验证，都取得了不错的效果。另外本文的方法可以有效的解决低频术语识别问题，为低频术语的识别提供新的思路。

关键词 URL；URL-Key；领域术语；低频术语；SVM；

Chinese Term Extraction Using URL-key

Lv Shuning[1]，Dong Zhian[2]

1. Software Engineering, Beijing University of Technology, Beijing 100124; 2.Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101

Abstract This paper presents a new approach for Chinese term extraction using URL-key. Taking the URL as a medium, with the help of known domain of URL-key, we can judge the domain which candidate terms belong to. First, with the help of domain URL classified artificially in the Internet, a method based on the variance is proposed to identify the domain URL-key and the dictionary of domain URL-key is built according to the frequency of URL-key appearing in various fields. Then, we use the pseudo related feedback to construct the URL-key vector of candidate domain terms. Finally, we apply SVM to extract terms. We conduct experiments on four different domains for Chinese term extraction. Experimental results indicate that the approach proposed in this paper is quiet effective. In addition, the proposed approach can effectively solve the recognition problem of low frequency terms, which provides a new way for the identification of low frequency terms.

随着时代的发展，互联网领域已经发生了巨大的变化，人们已经不再单纯的通过网络来获取数据，而是互联网数据的创造者。现如今已经进入了“大数据”时代，数据信息不仅规模大，而且错综复杂。新的理论、新的方法、新的概念等不断涌现，于此同时产生了大量新的领域术语，人工的构建领域术语不仅费时、费力，而且又不易更新，因此领域术语自动识别已经成为汉语自然语言处理重要的研究课题。

术语识别是基础性的研究工作，有助于领域词典的更新，领域本体的构建，以及句法分析的研究。术语识别研究通常分为两个步骤，候选术语提取和识别。候选术语提取可以看成是术语边界识别的问题。然而对于汉语来说，字符之间没有明确的切分边界，识别起来非常困难，文献[1]提出了利用特定词汇作为边界取代寻找候选术语与其上下文之间的特征关系，文献[2]通过观察选取候选术语的上下文边界信息，利用CRF对候选术语进行识别。有学者把术语识别的过程看成候选术语在不同领域语料中分布的过程。例如利用TF-IDF[3]、信息熵[4]等方法进行研究。还有学者更倾向于使用直接统计的方法，根据候选术语的频次、语义、上

国家自然科学基金项目（61671070）国家语委重点项目（ZDI135-53）

下文信息[5][6][7]来进行术语提取，而这些方法通常需要有一个大而精确的领域术语语料库作为基础知识进行学习。不幸的是目前尚未有公开的大规律领域术语语料库。

在本文的研究中，放弃使用领域语料库资源来进行识别，而是利用互联网中已标注的领域信息，提出一种新颖的利用Url-Key的领域术语提取方法。

URL 作为 Internet 上用来描述信息资源的字符串，近些年被学者们所关注，分别应用在 web 网页主题分类[8]、Query 分类[9][10][11]、广告关键词提取[12]等研究课题上。URL 主要由三部分组成：协议、域名、路径。域名通常由 2 部分构成：域名主体、域名后缀。对于 match.sports.sina.com.cn 这个域名来说，“match.sports.sina”为域名的主体词，而“com.cn”为该域名的后缀，而域名的主体词通常与信息资源的类别有关。而路径则是指信息资源具体存放的位置，人们在构建路径时，为了方便信息资源归类，通常会考虑将不同的资源放在不同的路径下，且为路径起的名称与资源的主题相关，如 football 的路径下应该放与足球相关的信息资源。

通过对 URL 的上述分析发现，人们在申请域名或者创建路径时，通常会思考与信息资源的相关性。虽然各网站的域名不同，但是相同主题下的域名主体词及路径引导词的使用可能相同。通常会集中使用一些领域性强的词语，如体育类中的“sport”、军事类中的“mil”、汽车类中的“auto”等。也就是说互联网中的域名主体词及路径引导词被赋予人们智慧的结晶。本文将 URL 中含有领域信息的域名主体词及路径引导词统称为**URL-Key**。

本文利用文献[2]提出的方法提取候选领域术语，将候选术语作为关键词串放入搜索引擎中进行检索，得到相关的反馈信息。与以往的研究不同，本文放弃了所有文本反馈信息，而是只使用反馈的 URL 作为桥梁，利用 Url-Key 的领域性来判断 URL 的领域性，从而判定候选候选术语的领域性。例如：候选领域术语为“凯美瑞”，搜索引擎返回的 URLs={“http://www.52**car**.net/”、“http://www.yi**cars**.com/”…}，由于 car 是汽车类 Url-Key 的关键词，那么 http://www.52**car**.net/和 http://www.yi**cars**.com/为汽车领域的 URL，所以很容易看出“凯美瑞”为汽车领域的领域术语。本文根据 URL-key 在不同领域中使用的差异性,对文献[13]中收集的领域 URL 进行领域 URL-Key 提取,构建领域 URL-Key 词表。再根据候选领域术语反馈的 URL 结果集，构建候选领域术语的 URL-Key 向量，最终利用 SVM 提取领域术语。

本文的主要贡献有3点，首先充分的借助互联网中的群体智慧，利用互联网中已存在的领域URL，提取领域URL-Key资源。其次首次提出了利用领域URL-Key进行领域词提取，从而解决了领域语料库资源稀缺问题，另外本文提出的方法可以快速、方便的移植到其它领域和语言上进行应用，为领域词提取提供了新的思路。最后本文的方法可以有效的解决低频领域术语提取问题。

1 领域术语识别

通常来说，一个句子是由领域术语和非领域术语组成，领域术语通常为实词，非领域术语通常为虚词、停用词或通用词等。非领域术语通常分布在领域术语的左右，构成了领域术语切分标记，利用切分标记，很容易把领域术语从句子中分离出来。

例如：*埃博拉病毒*是一种引发*埃博拉出血热*的*烈性传染病*。

句子中“埃博拉病毒”、“埃博拉出血热”、“烈性传染病”为健康医疗领域的领域术语，他被“是”、“一种”、“引发”、“的”切分标记所切割。再借助搜索引擎得到候选术语反馈的 URL 结果集。根据 URL 中含有 URL-key 的分数，构建候选领域术语的 URL-Key 特征向量，最终利用已知的 URL-Key 的领域性，对未知的候选领域词进行领域性判断，实现从候选领域术语的领域未知性到候选术语领域已知性的转变。

1.1 基于切分标记的候选领域词提取

利用文献[2]提出的 *TCE_DI*(Term Candidate Extraction –Delimiter Identification)。在领域语料库中使用切分标记库 *DList* 分割句子，将一个长句子切分成若干个小片段，在以词为单位进行组合提取候选领域词。如图 1 所示，给定句子 S= C1C2C3C4C5C6C7C8C9…Cn，其中 Ci 代表一个汉字，C1C2 构成词 Wj。句子中有两个切分标记 D1=C3、D2=C7C8，D1∈*DList*、D2∈ *DList*。*DList* 把句子切分成 3 个片段，TC1=C1C2、TC2=C4C5/C6、TC3=C9Cn，拼接后构成了候选领域词集合 TC={C1C2、C4C5、C6、C4C5C6、C9Cn}。

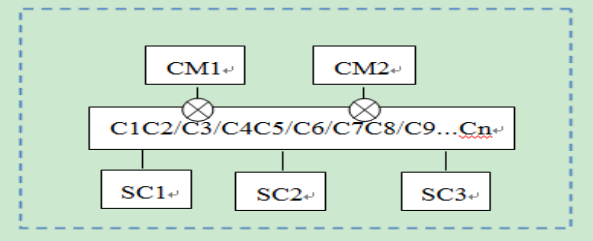

图 1 句子切分样例

Figure 1. The sample of sentence segment

1.2 基于 URL-Key 的领域词提取

本文利用方差的方法构建领域 URL-Key。利用搜索引擎伪反馈技术，得到相关候选术语的 URL 反馈信息结果。根据反馈 URL 集合中含有领域 URL-Key 的个数来构建候选领域术语的特征向量。最后利用 SVM 分类器判断是否为领域术语。

1.2.1 领域 URK-Key 构建

在互联网中很多资源都凝聚了人类的群体智慧，如分类网站，它们通常是经过人工整理分类后呈现给用户的，具有很高的可信度。文献[13]采集了 Yaboo、Google、Baidu 的中文分类网页目录,以 Yahoo 网页目录的前 2 层作为分类的标准，并将 Google 和百度网页目录全部映射到 Yahoo 的类别体系中，具体 URL 分布如表 1 所示。

表 1 URL 分布

Table 1 Distribution of URL

Topic	体育	汽车	军事	医疗
URL	1502	414	255	2353

为了得到更多的领域 Url-Key，需要找到更多的领域 URL。由于网络中存在着海量的 URL，且每天都在不断的增长，因此领域 Url 的收集将非常的困难。通过对文献[13]收集的 URL 分析发现，URL 在构建时词语使用存在着差异性，分布不均衡，特定词语使用的频次较高，如:“sport”在体育类 URL 中大量的出现，“car”在汽车类 URL 中大量的出现，“mil”在军事类 URL 中大量的出现，因此可以用根据 URL 是否含有 URL-Key 来判断 Url 的领域性。这样可以大大的缩减收集领域 Url 的时间，同时将收集领域 URL 的工作转换为提取领域 URL-key 的工作。

本文利用分割符 T={:，/,.,-}对文献[13]收集的 URL 进行切分，构成一系列的字符串，删除长度为 1 的字串、纯数字串和一些无意义的特殊字符序列如：http、www、com、cn 等。最终构建候选字符串集合为 CandidatekeySet。如：http://www.tennis.com/player/539/na-li/ 经过切分后得到 Candidatekey 为 tennis、player、na、li。根据 Candidatekey 在每个类别中分布的差异性，利用基于方差的方法构建 URL-Key，具体方法请参照文献[9]。

1.2.2 基于伪反馈的 URL-Key 特征向量构建

伪相关反馈是假设搜索引擎系统查询反馈的搜索结果排名越靠前与查询越相关[13]。本文基于以上假设，构建 *TC* 的 URL-Key 特征向量，首先利用 T 集合对伪反馈 URLs={url_1,url_2,…,url_l} 进行切分，构建一个 bag_i={$Candidatekey_{i1}$, $Candidatekey_{i2}$,…$Candidatekey_{ij}$}，目标类别为 C={c_1,c_2,…,c_n}，其中 i 为搜索引擎返回的前 i 条结果，而 j 为第 i 个 URL 切分后获取 Candidatekey 的个数，n 为类别个数。为了构建 *TC*

的领域 URL-Key 向量，需要计算 *TC* 伪反馈 URL 结果中每一类的 URL-Key 分数 Score(c_n|query)。

$$\mathrm{Count}(Url-key_n)=\begin{cases}1 & if \quad bag\text{中含有}\mathrm{URL-Key}\\ 0 & if \quad bag\text{中不含有}\mathrm{URL-Key}\end{cases} \tag{1}$$

利用公式 1 计算 *TC* 在各个类别中含有的个数。由于反馈结果中排名越靠前其与查询串越相关，则反馈的 URL 与 *TC* 的相关性与位置有一定的关系，也就是说 URL 中含有 URL-Key 且其位置越靠前 Score(c_n|*TC*)越高。本文将反馈结果的前 10 个查询结果看成权重相同的，后面结果随着排名的增加权重也逐渐降低，具体分布如公式（2）所示。

$$Pos(i)=\begin{cases}1 & (1\le i\le 10)\\ \dfrac{1}{\log(i+1)} & (i>10)\end{cases} \tag{2}$$

综上所述，查询串的 Score(c_n|*TC*)与反馈结果中含有的领域 URL-Key 有关，同时也与位置信息有关。其计算公式如公式（3）所示。

$$Score(c_n|TC)=\sum_{i=1}^{l} count(\mathrm{URL-Key})\times pos(i) \tag{3}$$

根据每一类含有URL-Key的分数值，最终构建出*TC*的特征向量{ $Score(c_1|TC)$ 、 $Score(c_2|TC)$ 、 $Score(c_n|TC)$ }。

2 实验与结果分析

2.1 实验数据

实验数据为 sogou 实验室开放的搜狐新闻数据，该数据收集了搜狐新闻 18 个频道不同数据，本文以体育、军事、汽车、医疗 4 个领域作为实验数据具体分别如表 2 所示

表 2 实验语料

Table2. Experimental data

Corpus	语料数据文档数	实验数据文档数
体育	419,768	100
汽车	29116	100
军事	13958	100
医疗	30790	100

为了解决领域词串低频提取问题，本文在不同领域文档中分别抽取前 100 篇文档做实验，并且这 100 篇文章涉及的内容各不相同，如，篮球、足球、乒乓球；汽车维修、汽车销售、汽车介绍…等等。因此词语分布分散，会出现大量低频词。如：润滑油出现 1 次、引擎盖出现 1 次、核动力出现 1 次、侦察机出现 1 次、感冒药出现 1 次、感觉神经功能障碍 1 次等，因此可以模拟出一个真实的领域低频词汇语料库。

2.2 评价指标

人工的统计文章中出现的所有的领域术语比较困难，因此人工标注经过 TCE_DI 处理后的候选领域术语。本文使用正确率作为实验的评价指标。

$$\mathrm{Pr}ecision=\frac{N(Correct)}{N(Detected)}*100\% \tag{4}$$

其中 N(Correct)表示正确检测出的领域术语，N(Detected)表示检测到的领域术语的总数。

2.3 实验与分析

2.3.1 领域 URl-Key 分析

领域 URL-Key 的提取是解决问题的关键，本文通过对领域 URL 的切分，利用方差计算公式计算每一个候选领域 URL-Key 的领域度，再根据领域度的大小进行排序，取 top100 作为候选领域 URL-Key。如表 3 所

表 3 部分领域 URL-Key

Table 3 Examples of URL-Key

File	Url-key
体育	sports、sport 、nba、espnstar、pingpang、guojizuqiu、basketaball、snooker、soccer、tennis、badminton、olympics、olympic、tiyu、tiyv、ty、zhongchao、cbachina、weiqi、guoneizuqiu、sportsbl、chess、saichang、nba、wangqiu、cba、volleyball
汽车	auto、car、xcar、vw、toyota、chinacars、Dongfeng、nissan、vehicles、audi、bitauto、new_cars、chery、newcar、peugeot、webcars、qiche、autohome、che
军事	mil、military、military、war、army、chinamil、armystar、junshi、js、milnews、tiexue、xinjunshi、
医疗	Hospital、beauty、yxy、eat、pharmacy、pharmacy、familydoctor、health、disease、zaojiao、jiankang、pathology、chinamedi

通过对领域 URL-Key 分析发现，领域 URL-Key 主要有以下特点：第一由英文单词或者英文缩写组成，如 sport、auto、car、hospital、nba、vw 等；第二为拼音或者拼音缩写，如 tiyu、che、js、zaojiao 等；第三为特殊网站域名，如：autohome、zhibo、tiexue 等。

为了得到准确的 Url-Key，根据领域 URL-Key 的上述特点，人工的过滤掉不相干的领域 URL-Key。最终得到 4 个类别，272 个领域 URL-Key，分布如表 4 所示。

表 4 领域 Url-key 个数分布

Table 4 Distribution of domain Url-key

Topic	体育	汽车	军事	医疗
Url-Key	69	74	55	74

2.3.2 候选领域术语

利用 TCE_DI 算法对 4 个候选领域语料进行切分，共提取候选术语 17123 个，其中体育领域候选术语 4681 个，汽车领域候选术语 3694 个，医疗领域候选术语 4491 个语，军事领域候选术语 3343 个。经过专家人工标注，最终分布如表 5 所示。其中体育领域有 1658 个，领域术语占总候选术语的 35.4%，有 3023 个非领域术语占总候选术语的 64.6%。

表 5 术语个数分布

Table 5 Number distribution of Corpus term

Corpus	候选领域术语个数	领域术语个数	非领域术语个数	领域率
体育	4681	1658	3023	35.4%
汽车	3694	1539	2155	41.6%
军事	3344	1196	2148	35.8%
医疗	4491	1888	2603	42.0%

由于 TCE_DI 算法将语料被切分若干个片段，根据每个片段中包含的词个数不同，将候选领域术语分为 1 元候选术语，2 元术语候选术语，和多元候选术语。具有分布如表 6 所示。从表中我们可以看出 2 元术语与多元术语中含有正确的领域术语个数比较多，这是因为有些词语单独出现没有实际的意义，但是和领域性强的词结合出现，却变成了领域词。如：

表 6 术语分布表

Table 6　Distribution of domain term

Corpus	词形	数量	正确数	正确率
体育	1 元候选术语	2406	682	28.3%
	2 元候选术语	1481	702	47.4%
	N 元候选术语	794	274	34.5%
汽车	1 元候选术语	1509	364	24.1%
	2 元候选术语	1680	827	49.2%
	N 元候选术语	505	348	68.9%
军事	1 元候选术语	2292	541	23.6%
	2 元候选术语	683	391	57.2%
	N 元候选术语	369	266	72.0%
医疗	1 元候选术语	2899	989	34.1%
	2 元候选术语	1037	557	53.7%
	N 元候选术语	555	343	61.8%

2.3.3 术语识别

本文把领域术语识别任务看成一个二分类的任务，利用 SVM 对于候选术语分类。分别随机抽取各领域 80%的正确的候选术语和错误的候选术语作为训练语料，20%正确的候选术语和错误的候选术语作为测试语料。其中体育领域随机抽取 3734 个候选术语作为训练语料，947 个候选术语作为测试语料，汽车领域

随机抽取 2882 个候选术语作为测试语料，812 个候选术语作为测试语料，医疗领域随机抽取 3582 个候选术语作为测试语料，909 个候选术语作为测试语料，军事领域随机抽取 2564 个候选术语作为测试语料，780 个候选术语作为测试语料。具体分布，如表 7 表 8 所示。

表 7 训练数据

Table 7　Train data

Corpus	体育领域		汽车领域		军事领域		医疗领域	
词语类型	训练术语	训练非术语	训练术语	训练非术语	训练术语	训练非术语	训练术语	训练非术语
1 元词	544	1377	287	915	431	1400	789	1526
2 元词	560	621	638	642	211	232	443	383
N 元词	217	415	277	123	210	80	273	168
总计	1321	2413	1202	1680	852	1712	1505	2077

表 8 测试数据

Table 8　Test data

Corpus	体育领域		汽车领域		军事领域		医疗领域	
词语类型	测试术语	测试非术语	测试术语	测试非术语	测试术语	测试非术语	测试术语	测试非术语
1 元词	138	347	77	230	110	351	200	384
2 元词	142	158	189	211	180	60	114	97
N 元词	57	105	71	34	56	23	70	44
总计	337	610	337	475	346	434	384	525

本文利用搜狗搜索引擎，把候选领域术语当作检索词条进行检索，爬取前 100 条 URL 作为数据的实验集合，根据 URL 实验集合中含有 URL-Key 的个数来构建候选领域术语的 URL-Key 特征向量。为了比较 URL-Key 与 URL 的方法，本文将领域 URL 直接与爬取的 URL 实验集合进行匹配，构建 URL 特征向量。实现结果如图 2 所示。从图中可以看出利用领域 URL-Key 的方法要好于利用 URL 匹配的方法，这是因为搜索引擎反馈的 URL 中不可准确的与人工搜集的 URL 完全的匹配。据统计，每返回 100 条 URL 时，有大约 20%的 URL 能够进行 URL 匹配，大约有 40%的 URL 能够进行领域 URL-Key 的匹配。因此利用 URL-Key 的方法可以解决互联中海量 URL 每天不断的增长的问题。

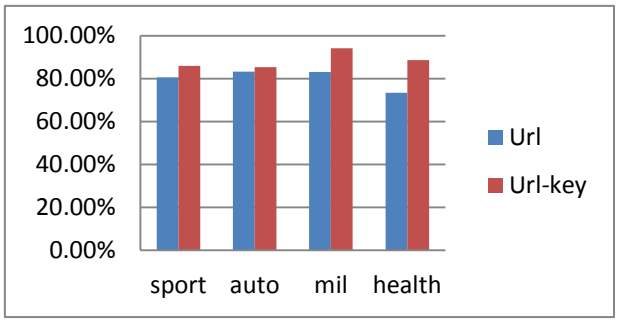

图 2URL 与 URL-key 对比实验结果

Figure 2. Contrast result of URL matching and URL-key

在 URL-Key 匹配的过程中发现检索的 URL 中存在着复合领域字串 attatclvolleyhball，文献[10]以字符为单位利用 all-grams 把 volleyball 提取出来，而本文无需对复合的领域词进行处理，只需要通过控制候选术语爬取 URL 的数量，弥补复合词带来的缺陷。是否爬取的 URL 越多实验结果越准确呢？本文对爬取不同数量的 URL 个数进行了分析，如图 3 所示，不是 URL 的数量越多实验的结果越准确。当 URL 取前 70 个时实验结果最好，这是由于在搜索引擎中，网页越靠后，网页与搜索词之间的相关性就越疏远。

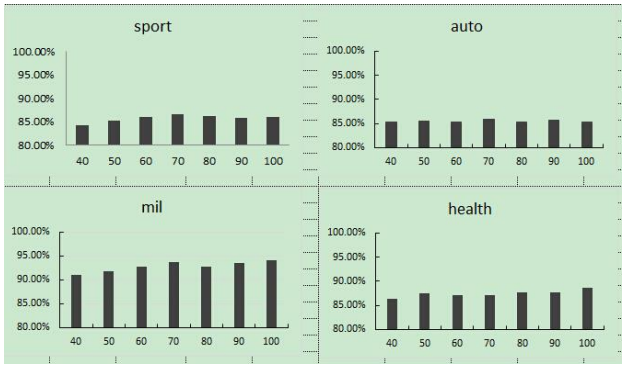

图 3 不同 URL 个数的对比实验结果

Figure 3. Contrast result of different URL number

利用 URL-Key 作为特征，对每个候选术语爬取前 70 个 URL 作为实验结果集。实验结果如表 7 所示，体育领域评价正确率 86.06%，汽车领域的评价正确率为 87.77%，军事领域的平均正确率为 93.66%，健康医疗领域的平均正确率为 87.17%。从表 9 可以看出，在各领域中 2 元词的正确率要低于 1 元词和 N 元词，这是因为 2 元词中会有一些单字组成的词语，这些词语，通常为领域术语的一部分，自己本身却构不成领域术语，但在检索返回的 URL 的实验集合中，却出现很高的 URL-Key 的匹配率，例如：“款/福克斯”“式/战机”“性/皮炎”这些不是领域词，但是这些词语 URL-Key 的匹配率都超过了 40%，这是因为搜索引擎会根据用户提供的检索词进行自动的扩展检索，已“款福克斯”为检索词，返回的结果中都是关于“新款福克斯”的信息，或者是“福克斯”的信息，所以这种情况会影响 2 元词的准确率。

表 9 实验结果正确率

Table 9. Experiment results

Corpus	体育领域	汽车领域	军事领域	医疗领域
1 元词	0.9079	0.8679	0.9239	0.8672
2 元词	0.8290	0.8229	0.9126	0.8373
N 元词	0.8448	0.8823	0.9733	0.9107
平均值	0.8606	0.8777	0.9366	0.8717

本方法与文献[14]提出的 TV_LinkA 方法进行对比，由于文献评价指标为 Top-N，所以本文要根据各测试语料中含有的领域术语个数进行判断，体育领域取 top-337，汽车领域取 top-337，军事领域取 top-384，健康领域取 top-346 作为评价指标。而本文认为 SVM 在测试语料中识别正确术语的正确率与对比实验的正确率可以直接对比。实验结果如表 10 所示。

表 10 对比实验结果

Table 10. Contrast results

Corpus	体育[top337]	汽车[top337]	军事[top384]	医疗[top346]
本文方法	86.67%	87.77%	93.66%	87.17%
TV_LinkA	66.40%	65.77%	70.63%	69.88%

从表 10 可以看出 TV_LinkA 的实验结果没有论文中描述的高，分析主要存在两种原因，第一领域语料不同，文献中应用的为 IT、Leagl 类语料，本文应用的为体育、汽车、军事、健康类语料，语料不同质，有可能导致结果的差异。第二原因，语料规模的大小，文献中应用的 IT 领域的数据集 6.64M。为了构建真实的低频数据，体育领域数据集为 266KB、汽车领域数据集为 219KB、军事领域数据集为 205KB、健康领域数据集为 196KB。而文献应用 HIST 算法，根据候选术语，提取领域句子，由于候选术语出现的频次低，所以提取的领域句子的数量会变少，将影响 HIST 算法的结果，因此语料规模的大小是影响 TV_LinkA 方法的最主要原因。

而本文的方法无需考虑实验数据规模问题，实验方法非常的稳定。无论对低频的术语或高频的术语本方法都非常适用。

2.3.4 低频术语识别

把候选术语集中出现次数为 1 次的称为低频术语，各领域的具体分布如表 11 所示。从表 9 可以看出本文的实验的语料低频率是非常的高，其中，体育领域语料的低频术语 3354 个，低频率为 71.65%，汽车领域的低频术语 2674 个，低频率为 72.39%，军事领域的低频术语 2239 个，低频率为 66.96%，医疗健康领域的低频术语 3059 个，低频率为 68.11%。

表 11 各领域低频分布表

Table 11. Distribution count of every domain

Corpus	候选术语类型	数量	低频术语	低频率
体育	1 元候选术语	2406	1578	65.58%
	2 元候选术语	1481	1102	74.41%
	N 元候选术语	794	674	84.89%
汽车	1 元候选术语	1509	989	74.82%
	2 元候选术语	1680	1257	74.82%
	N 元候选术语	505	428	84.75%
军事	1 元候选术语	2292	1368	59.68%
	2 元候选术语	683	555	81.25%
	N 元候选术语	369	316	97.83%
医疗	1 元候选术语	2899	1652	56.98%
	2 元候选术语	1037	904	87.17%
	N 元候选术语	555	503	90.63%

在以往的研究中低频术语的识别率非常的差，本文提出的方法可以很好的解决这个问题。在 4 个领域，3448 个测试候选术语中，共有 2422 个低频候选术语，占总共测试集的 70.25%。正确识别出 2128 个候选术语，其中 638 个术语，1490 非术语。错误识别 294 候选术语，其中 70 个术语，224 个非术语。实验结果如表 12 所示。

表 12 低频实验结果

Table 12. Low frequency experimental results

结果	候选术语	术语	非术语
正确识别个数	2128	638	1490
错误识别个数	294	70	224
正确识别率	87.86%	89.03%	84.96%

对于普通方法，以单篇文章语料抽取领域术语时，通常舍弃频次为 1 的候选术语，这样的做法将导致大量的术语不能够被准确的识别。本文方法得到了一个令人兴奋的结果，低频词语识别准确率达到了 87.86%，真实准确率达到了 89.03%，很好的解决了低频术语识别问题。此外新词识别的研究中，新词在最初的语料中出现的频次也是很低的，因此可以推断该方法对于新词识别同样有效。

3 总结

本文利用互联网中凝聚人们智慧的 URL，提取领域 URL-Key。首次提出了利用 URL-Key 进行领域词提取的方法。有效的解决了由网站爆炸式增长带来的领域 URL 不匹配问题，同时也解决了领域低频术语的识别问题。

首先利用方差提取领域 URL-Key，构建领域 URL-Key 词表，再利用伪反馈技术，构建候选领域词的 URL-Key 特征向量，使用 SVM 对候选领域术语进行提取，最终得到领域术语。根据领域术语的个数，研究了低频领域术语的识别率，实验表明本文方法在低频术语识别的准确率为 87.86%。提出的利用 URL-Key 提取领域词的方法，在一定程度上解决了领域词提取的问题，同时也对低频术语的识别提供了新的思路。

本文提出的方法仍有需要改进的地方，例如领域 URL-key 构建的精确度，SVM 选取更多特征，在更多领域进行实验。这些将在未来的工作中进一步研究。

参考文献

[1] Ji L, Sum M, Lu Q, et al. Chinese terminology extraction using window-based contextual information[M]//Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2007: 62-74.

[2] Yang Y, Lu Q, Zhao T. Chinese Term Extraction Based on Delimiters[C]//LREC. 2008.

[3] Salton, G., and McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

[4] Chang J S. Domain specific word extraction from hierarchical Web documents: A first step toward building lexicon trees from Web corpora[C]//Proceedings of the Fourth SIGHAN Workshop on Chinese Language Learning. 2005: 64-71.

[5] Ji L, Sum M, Lu Q, et al. Chinese terminology extraction using window-based contextual information[M]//Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2007: 62-74.

[6] Sornlertlamvanich V, Potipiti T, Charoenporn T. Automatic corpus-based Thai word extraction with the C4. 5 learning algorithm[C]//Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2000: 802-807.

[7] 木合亚提·尼亚孜别克,古力沙吾利·塔里甫. 哈萨克语 IT 领域术语识别研究与实现[J]. 中文信息学报,2016,(03):68-73.

[8] Qi X, Davison B D. Web page classification: Features and algorithms[J]. ACM Computing Surveys (CSUR), 2009, 41(2): 12.

[9] 李雪伟,吕学强,董志安,刘克会. 利用 URL-Key 进行查询分类[J]. 北京大学学报(自然科学版),2015,(02):220-226.

[10] Baykan, Eda, et al. "Purely URL-based Topic Classification." Proceedings of International Conference on World Wide Web3.10(2009):1689 - 1694.

[11] Shen D, Sun J T, Yang Q, et al. Building bridges for web query classification[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 131-138.

[12] Raju S, Udupa R. Extracting advertising keywords from URL strings[C]//Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012: 587-588.

[13] 张宇,宋巍,刘挺,李生. 基于 URL 主题的查询分类方法[J]. 计算机研究与发展,2012,(06):1298-1305.

[14] Yang Y, Lu Q, Zhao T. Chinese term extraction using minimal resources[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 1033-10